УДК 004.82

МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ДАННЫХ В СЕМАНТИЧЕСКОЙ СЕТИ

Егошин А.В. (Университет ИТМО)
Научный руководитель – старший преподаватель факультета программной инженерии и компьютерной техники Цопа Е.А.
(Университет ИТМО)

Введение. В современном мире когнитивное восприятие информации является важным элементом информационных систем. Организация информации в виде семантических сетей или тезаурусов позволяет структурировать смысловые концепты в иерархию. В работе [1] была предложена схема хранения связей концептов и данных, их выражающих в общую структуру, которая организована по образу и подобию современных базы данных. При этом к семантической сети можно использовать запросы для выборки данных с использованием родо-видовых и композиционных связей в существующей структуре. Использование семантической сети как структуры базы данных позволяет решать такие задачи выборки информации из текстов на естественных языках, снятие полисемии и омонимии для заданных предметных областей.

Основная часть. Хранение данных в семантической сети, разрабатываемой коллективом авторов в рамках исследований в Университете ИТМО и сторонних организациях организованно в виде сегмента разделяемой (shared) памяти, который сохраняется в файле средствами операционной системы. Данные о смысловых концептах и их экземплярах, словоформах и поисковых индексах хранятся в виде структур boost::multiindex. Над хранимыми данными организована модель конкурентного доступа для транзакций MVCC, которая позволяет обрабатывать запросы в многопроцессорном и многомашинном режиме. Верхнеуровневое API использует модель транзакций и организует многопользовательский сервер, который обрабатывает запросы прикладных пользовательских программ, работающих с текстом на естественном языке.

Существующая модель хранения обладает рядом недостатков, обусловленных выбранными архитектурными решениями. Например, структуры boost::multiindex прекрасно справляются с запросами к данным, автоматически строят и обновляют поисковые индексы, но размещение их в разделяемой памяти ведет к записи состояний блокировок узлов в файл, и если сервер семантической сети аварийно завершился, блокировки могут остаться в захваченном состоянии внутри файла данных. Кроме того, важным моментом является повышение скорости запросов к структуре и данным семантической сети.

На основании выявленных недостатков, в рамках работы был осуществлен пересмотр модели хранения данных и структуры семантической сети. Вместо использовавшейся унифицированной схемы были реализованы отдельные структуры данных для организации хранения узлов семантической сети различных типов и отношений между ними. Предложенная модель хранения данных ориентирована на ускорение операций извлечения данных на основе комплексных запросов, требующих ресурсоемких обходов графа семантической сети.

В системе продолжает использоваться сегмент разделяемой памяти для кэширования часто используемых данных и организации доступа служебных процессов с использованием разделяемой памяти, но этот сегмент не отображается на файл в операционной системе, что позволяет избежать проблем с сохранением блокировок в файлах и повышает производительность системы за счет уменьшения числа обращений к файловой системе.

Предложенные архитектурные изменения были реализованы в рамках существующего программного модуля на языке программирования C++. По итогам реализации было проведено функциональное и нагрузочное тестирование, показавшее более высокую эффективность модернизированной системы по сравнению с исходной моделью хранения данных в задачах обработки текста на естественном языке.

Выводы. В рамках работы была полностью пересмотрена модель хранения структуры и данных семантической сети. Данная модель была реализована в составе программного модуля на языке C++ и протестирована на примере извлечения синонимов для исходного набора понятий из семантической сети, построенной на основе тезауруса Wiktionary и показала свою эффективность для задач систем обработки естественного языка.

Список использованных источников:

1. Клименков С.В., Николаев В.В., Харитонова А.Е., Гаврилов А.В., Письмак А.Е., Покид А.В. Применение семантической сети для хранения слабоструктурированных данных // Инженерный вестник Дона [электронный журнал] - 2020. - № 2(62). - С. 27
2. Письмак А.Е., Харитонова А.Е., Цопа Е.А., Клименков С.В. Метод автоматического формирования семантической сети из слабоструктурированных источников. Программные продукты и системы. 2016. № 3. С. 74-78.
3. Tsopa E., Mileshin A.A., Slapoguzov A., Klimenkov S. Sentiment Analysis of Russian Text Using the Semantic Network. XIII Международная научно-практическая конференция молодых ученых «Программная инженерия и компьютерная техника» (-2021), MICSECS 2021 SAINT-PETERSBURG The Majorov International Conference on Software Engineering and Computer Systems Международная конференция. (Санкт-Петербург, СПб, 1-3декабря 2021г.). 2021. Vol. 2. pp. 100.